ISSUES ON THE ACQUISITION OF ITALIAN COMPLEX NOMINAL FROM TEXT CORPORA: A COMPUTATIONAL APPROACH COMBINING SYNTACTIC AND SEMANTIC INFORMATION

Valeria Quochi

Introduction

Complex Nominals, as a subset of Multi Word Expressions (or collocations)[1], represent a serious problem both for any theory of the lexicon and for lexicography, whether traditional or computational. The main reason for this is that they appear not to respect the traditional distinction between syntax and the lexicon. In particular, Italian Complex Nominals (ICNs hereafter) show a great variability in degree of lexicalisation and discontinuity: the range goes from fully lexicalised ones (like *macchina da cucire* 'sewing machine') to productive and regular ones (like *scatola di vetro* 'glass box'), which are regarded, in traditional grammars of Italian, as different types of complements. Because the N+PP type is structurally regular and therefore does not pose great problems during syntactic analysis, it has been often ignored, especially in the computational tradition. However, when semantic interpretation is taken into account, things get more complicated. In fact, the same syntactic pattern may receive different semantic interpretations. This depends both on the lexical items involved and on the linguistic and extra-linguistic context of utterance. An account of CNs within the lexicon, while necessary, is extremely problematic, thus highlighting the inadequacy of many well-established theories, recent models and computational systems. The problem of the representation of CNs is serious from both a theoretical and a practical perspective, not only because CNs are difficult to classify and identify in texts, but also because it is not clear yet what information must be encoded and how. At present, completely fixed ICNs appear to be less problematic: very often they deserve a full entry in the lexicon, and recent statistical

[1] *Multi-Word Espressions* is a term born and principally used within the computational community, whereas *collocation* is the term more frequently used in lexicography and lexicology.

approaches allow for acquisition of fixed Multi Words from texts in an almost automatic way, with few false positives. The more "productive" ones, instead, still present problems: it is precisely the identification of their regularities that aroves difficult, given the available grammatical prameworks. By regular ICNs we mean well-established and productive syntactic-semantic paradigms: i.e. syntactic expressions that obey the standard rules of the grammar and that instantiate some semantic paradigms (as an example consider the "material complement" *scatola di cartone* 'cardboard box'). Such expressions, which will be the focus of the present contribution, need to be recognized as some kind of unit once we want to assign them a semantic interpretation, which is partially idiosyncratic, or unpredictable on purely syntactic basis. Moreover, it is noteworthy that these patterns may function as the basis for lexicalisation: *occhio di vetro* (lit. 'eye of glass') is an example of a lexicalised CN, that is an expression, based on the made of pattern, wich has been conventionalised, thus acquiring other meaning components such as that it replaces a real eye in a human being.

This paper will not give a solution to the problem. The aim of the experiment we have been conducting, and which builds on a previous pilot study (Quochi 2004), is to identify the highest possible number of productive syntactic-semantic patterns of ICN formation, and to make explicit the particular semantic relation that exists between the head of the phrase and its modifier. Recently, the implicit semantic relation underlying CNs has been considered as the most important information necessary for the its interpretation, especially of productive ones. Therefore, to identify the relation underlying CNs is particularly relevant for computational treatment of ICNs and a fundamental feature that must be represented in a lexicon. The main problem is how to formalise this relation: Lexical Functions- that are largely used in formal lexicography because they have a strong descriptive power, work well for predicative elements, but they do not seem to be appropriate for non-predicative ones, such as pure nouns. Therefore, we have based our experiment on a different theory of the lexicon, namely the Generative Lexicon Theory, that provides for a structured representation of the internal semantics of lexical items.

Italian Complex Nominals

In literature, Complex Nominal expressions generally include compound nouns and unpredictable adjective-noun pairs (Levi 1978). However, if we consider this issue under a cross-language perspective, we observe that many

English compounds, for example, must be translated into an expression consisting of a complex nominal group in a romance language like Italian: usually the translation equivalents of noun compounds are complex syntactic expressions consisting of a noun modified by a prepositional phrase (N+PP hereafter), where the preposition is *a* ('at'), *di* ('of') and *da* ('from/by'). This is one of the facts that led us to consider such N+PP expressions as CNs.

Italian CN formation mainly exploits post-modification, as we can see in the examples below; we find different structural types of ICNs, among which:

- Actual compound nouns, (N+N): *nave scuola* 'school-ship', *capostazione* lit. 'station master';
- Nouns modified by non-predicative adjectives (N+Adj_{npred}): *coltello elettrico* 'electric knife';
- Nouns modified by a PP (N+PP): *coltello da pane* 'bread knife'[2].

Compound nouns in Italian are not as frequent as in languages like English or German, and they are for the most part lexicalised, often foreign calques. N+Adj CNs are quite common but not very productive.

On the contrary, the N+PP type is both frequent and highly productive, and in fact it is the most frequently exploited in translation, especially in technical domains where there is a constant need to produce translations of new terms (see Petrocelli 1992). Nevertheless, ICNs of the N+PP type present particular difficulties because they are, in general, both syntactically regular and semantically (almost) transparent, although they show a great variability. The expressions that instantiate this syntactic type distribute all along the compositionality continuum: some N+PPs are fully lexicalised, fixed or idiomatic, others are regular and productive - though they sometimes show morpho-syntactic anomalies. All this brings into play the internal lexical semantic structure of nouns, esp. bare nouns, and the problem of the polysemy of prepositions.

In the following we will describe some morphological, syntactic and semantic characteristics of N+PP ICNs[3].

[2] See also Voghera (2004: 62-63) for a detailed list of Italian CN types.

[3] For the sick of simplicity, because the focus of the present contribution is exclusively the N+PP type, hereafter we will use the term ICNs to refer only to the N+PP type. The reader must bear in mind, therefore, that the present analysis does not fit other structural types of ICNs.

Morphosyntactic Characteristics of Italian CNs

At the morphological level of description, ICNs do not show particular anomalies. These CNs can be seen as consisting of a noun modified by another noun introduced by a simple preposition. The prepositions that can occur in N+PP complex nominals are *di*, *da* and *a*[4]. Often the modifier can occur either in the singular or in the plural form, but the choice seems not to be predictable, rather it seems a lexical choice, as shown in ex.1.

Ex. (1)
scatola da scarpe vs. **scatola da scarpa*

At the syntactic level of description they often show some anomalies with respect to "regular" syntactic expressions, that can be used as clues to isolate ICNs in text. However, to define a CN with respect to a structurally similar expression is not an easy task. Anomalies cannot be employed as rules to identify and generate all and only possible/existing ICNs, because they represent tendencies and exceptions are numerous.

First of all, the modifier noun generally occurs without determiner:

Ex. (2)
Scatola da scarpe 'shoe box' vs. *scatola dalle scarpe*
'box from shoes'.

Moreover, no lexical item can normally intervene between the head N and the PP:

Ex. (3)
Un bicchiere da vino ('a wine glass') but ??*un bicchiere buono da vin*o ('a wine glass made of good quality glass');
Una carta di credito (' a credit card'), but **una carta blu di credito* ('a credit blue card');
Una barca a vela ('a sailing boat'[5]), but *?una barca bianca a vela* ('a saling white boat').

In the examples above a modifier of the head noun is more felicitously in-

[4] Probably also N+PP expressions with different prepositions may be considered CNs, but at present other prepositions are not taken into consideration. (see also Giovanardi 2004: 584).
[5] NB: in AE the equivalent is *sail boat* or *sailboat*, which is parallel to the Italian ICN, differently from what happens in BE.

serted either in pre-nominal position (*un buon bicchiere da vino*, 'a good wine glass') or after the PP that is it must modify the whole expression (*una barca a vela bianca*, 'a white sailing boat').

Semantic Aspects of Italian Complex Nominals

Italian Complex Nominals share many of the semantic characteristics of Noun Compounds, as they have been described in the extensive literature, especially on the English language[6]. The main feature of CNs, is that they form a conceptual unit, whether permanent or temporary, that is they denote some kind of entity (Downing 1977).

Like English noun compounds, Italian CNs may undergo lexical semantic processes such as semantic specialisation and lexicalisation (for example *carta di credito*, 'credit card'). Moreover, like endocentric compounds, CNs are hyponyms of their semantic head, which in the case of ICNs is always the leftmost noun.

All the above-mentioned types of information are extremely useful for the interpretation of CNs in different contexts, and for the understanding of texts; and because they do not seem to be fully predictable, they certainly deserve an account in the lexicon.

The interpretation of CNs, however, has been said to depend crucially on the ability to retrieve the appropriate semantic relation that holds between its two component nouns. The general heuristic for the interpretation of CNs is to link the meaning of the component nouns with the appropriate semantic relation. The problem with noun compounds is that this relation is completely implicit. In ICNs, instead, the two nouns are syntactically linked by a preposition, which has been considered as a morpho-syntactic mark of the underlying semantic relation (Johnston & Busa 1999). Prepositions, however, are known to be highly polysemous elements (see Weinrich 1977), and their interpretation can only be assigned in context.

Despite the high degree of variation and the presence of a preposition, the semantic relations between the constituents of ICNs appear to be the result of the interaction of the semantics of both the head and the modifier noun, and there appear to be more systematic paradigms than one would

[6] See Levi 1978, Warren 1978, as important examples of works dedicated to noun compounds in English; about word formation in Italian see, in particular, Grossmann and Rainer 2004.

think[7]. For this reason, our interest lies not so much in fixed, highly lexicalised CNs, but especially in those CNs that appear to be quite regular both at the syntactic and at the semantic level, but which, nevertheless, pose problems if they are not recognized as units at some level of analysis[8]. The present investigation attempts to discover semantic paradigms of ICNs, in order to make the assigning of a semantic relation to ICN tokens easier and (semi-) automatic. The experiment has been carried out on corpus data, starting from a hypothesis of cross-language semantic interpretation of CNs made within the generative lexicon framework (Johnston and Busa 1999).

The Generative Lexicon Theory[9]

The following sections contain a description of the fundamentals of the theory of Generative Lexicon (GL hereafter) and its basic architecture, with particular attention to those aspects that concern the internal structure and the treatment of nominal elements (Pustejovsky 1995, Busa 1996).

GL theory aims at modelling and formalizing the lexical knowledge of native speakers of a language, taking into consideration the immediate linguistic context of single words, in order to account for the creativity of language use. The lexicon, according to Pustejovsky, must be considered an essential component of linguistic knowledge, and not a mere repository of all idiosyncrasies, as it has been seen in traditional generative (chomskian) models.

Specifically, the representation of lexical items is taken to be an important part of the composition rules that generate the set of all possible interpretation of words in context. Among the objectives of this model is the ability to account for all different meanings that words assume in context, without listing all of them in the lexicon, which would be practically impossible as well as theoretically anti-economic. Within the Generative Lexicon, word senses are no longer considered as atomic units of meaning, but as entities that bear

[7] In more traditional accounts, especially in those that treat CNs as collocations, this has not always been recognized, and the modifier (or collocate) has been seen as selected by the head (or base).

[8] They create problems, for example, in PP attachment during parsing; or in IR they often need to be searched as bound elements, instead of as separate words.

[9] The content of this section is basically a synthesis of the works of Pustejovsky (1995) and Busa (1996).

an articulated internal semantic structure and that are encoded at different levels of representation. The model, moreover, is seen as a system of various semantic relations among concepts/ senses. (Busa 1996: 44). In GL a lemma is represented as a meta-entry, i.e. an entry that subsumes all its different but related senses, a structure that permits encoding the regularities in semantic behaviour that become evident during the processes of composition and combination in context.

The fundamental assumption lying behind this theory is that lexical knowledge is to be kept distinct from world knowledge. The problem is that non-linguistic knowledge certainly contributes to lexical meaning; however, the language system has only a partial access to such information: the senses do not entirely reflect the deep conceptual structures of our cognitive system, but encode mainly those aspects that have some influence on the grammatical behaviour of lexical items (Busa 1996: 45). Non-linguistic knowledge allows for the comprehension and establishment of the semantic types that are fundamental for the linguistic system and the relations among them. This way it gives to the language system a basic conceptual structure (an ontology) that remains, nevertheless, distinct from it. A theory of this kind presupposes the existence of various levels of semantic analysis, and lexical semantics is just one of them. All such levels give independent interpretations, which contribute to the construction of the global meaning of a discourse, once adequately interrelated.

The GL Model

Word senses – and the concepts that these express- are related to other senses within the lexicon by means of a network of explicitly defined links and are combined in context by means of a set of generative mechanisms that make use of the semantic information given by the lexicon. The senses/concepts have an internal semantic structure that is in itself relational; the links among different senses are provided by elements of meaning captured within the Qualia Structure (see below).

GL is organized into four representational levels: Argument Structure, Event Structure, Qualia Structure, and Lexical Inheritance Structure. The Qualia Structure (QS hereafter) is the most original idea presented within this theory, and the one that is directly involved in the encoding of the semantics of nominals. For this reason, we will not discuss the other levels in the present paper and will concentrate on the Qualia Structure alone. For details the reader may refer to Pustejovsky (1995).

The Qualia Structure

QS is the level where logical arguments and events are linked through relations that explain and make explicit the meaning of lexemes. These relations are expressed under four Qualia Roles, which Pustejovsky calls "generative factors". According the GL theory, the Qualia roles fulfil the task of guiding the speaker's comprehension of the objects and relations in the world.

The qualia structure ultimately guarantees the internal cohesion of the whole lexicon. This is the level that provides a structured representation of semantic information describing the relational meaning of senses. It is organised in four roles, which specify four essential aspects of meaning: the Formal Role specifies what a sense denotes; the Constitutive Role specifies what an entity is made of, its inherent constitution; the Agentive Role specifies how the entity has come into being, who has created it or how it has been made; the Telic Role specifies the function.

There are two general principles underlying qualia roles: first, every lexical category expresses a qualia structure; second, not every sense has a specified value for each qualia role. The first principle guarantees a uniform semantic representation across all lemmas, whereas the second principle allows the lexicographer to consider qualia roles as specifiable according to the particular characteristics of each semantic class. The relational view of lexical meaning of the GL, thus, takes into consideration aspects of meaning that have been almost always ignored by traditional formal accounts of lexical semantics.

Compositional treatment of CNs within GL

Johnston and Busa (1999) propose a compositional analysis of English noun compounds and their equivalent N+PP Italian CNs based on the Qualia representation of lexical items and on the generative mechanisms provided by the generative lexicon theory. The Qualia representation provides nominal elements with a relational structure through which it is possible to describe how a head noun can be modified by another noun, making use of the same generative mechanisms that play a role in the interpretation of sentences, i.e. type coercion and co-composition. According to this approach the major difficulty lies in the lack of a systematic method for the retrieval of the semantic relations underlying CNs. Once the relation has been identified, it must be expressed in the qualia structure by specifying the semantic content of the modifier in the appropriate role of the head noun. A compound noun like *bread knife*, thus, can

be formalised in a qualia structure that will inherit from the head noun all information relative to the formal role and to its typical function; the modifier, instead, will specify the object of the typical activity of the head noun.

Ex. (4)

Bread Knife
Formal: isa: Instrument (inherited from knife)
Telic: 'to cut' (inherited from knife)
Object of cut: bread.

The Italian equivalent would be encoded in a parallel way:

Ex. (5)

Coltello da pane
Formal: isa: Strumento (inherited from *coltello*)
Telic: 'tagliare' (inherited from *coltello*)
Object of 'tagliare': pane.

It is worthy of note that in this approach the preposition would appear only in the citation form of the lemma, and not in the entry. This kind of representation, moreover, would allow for an automatic linking of translation equivalents, in that it would suffice to find two CNs with a parallel qualia structure.

It must be observed, however, that not all CNs can be easily formalised, this way, because it is not often clear in terms of qualia structure, what relation exists between the two main elements; a limit of the Qualia Structure (QS) becomes immediately evident: the four roles are too general, whereas more specific relations seem to be needed.

In the case of Italian N+PP CNs, additionally, there is a problem with the selection of the appropriate preposition, when generating CNs. According to Johnston and Busa (1999), the prepositions in N+PP (*di*, *da* e *a*) must be considered like bound morphemes, associated with one qualia role each: *di* would be associated with the agentive role, *da* would be associated with the telic role, and *a* would be associated with the constitutive role.

A corpus investigation has revealed that such generalisations do not hold: only the preposition *da* seems to be systematically associated with a telic relation, the others, especially *di*, are much more variable. Additionally, ambiguity between roles is very high. A CN like *fucile a pallettoni* ('shot gun') is ambiguous between a constitutive and a telic relation: the shots can be seen as the objects of the typical activity of the gun, or as a part of the gun itself.

Ex. (6)

fucile a pallettoni 'shot gun'
Formal: isa: Instrument (inherited from *fucile*)
Telic: *sparare* 'shoot' (inherited from *fucile*)
Object of *sparare*: *pallettoni* 'shots'

Or

Formal: isa: Instrument (inherited from *fucile*)
Constitutive: *pallettoni*

In such a case, it is difficult to decide which one is the best representation. Probably, both interpretations are plausible, and somehow both are made available by the ICN, but QS does not allow both "senses" to be encoded simultaneously.

Qualia Structure then, does not specify all possible combinations of a lexical item; it only claims to contain all relevant information that potentially allow for the generation of senses in context. Therefore, for ICNs that are not fully compositional, specific interpretation rules or representations seem to be needed. Regardless of the specific theoretical or representational approach, the covert semantic relation needs to be made explicit.

Given that the identification of the correct semantic relation is crucial for the interpretation /representation of ICNs, and that the preposition is not a good clue, not even to identify the relevant qualia role involved, we propose to look for syntactic-semantic paradigms of CN formation, wich rely on information about the semantic type of both the head and the modifier noun, and on the syntactic structural type instantiated.

The experiment: establishing semantic paradigms

The Data[10]

Our investigation is based on corpus data, obtained through two main steps. The first step consisted of a set of syntactic rules aimed at the extraction of

[10] The first steps of the experiment, the tuning of the input corpus, the rules for the extraction of ICN candidates have been designed and written by Raquel Marchi at ILC (CNR- Pisa). The same person has followed the extraction of the relevant data annotated with syntactic information. In this paper we give a brief summary of the whole process. A detailed description of these phases is given in a draft paper (contact: armarchi@gmail.com).

N+PP structures satisfying certain morpho-syntactic conditions. The second step consisted of augmenting the dataset annotated with the relevant syntactic information automatically extracted in the previous step with information on the semantic class of nouns.

The extraction of the dataset that served the present investigation was made from a 3-million-word corpus, balanced and representative of contemporary Italian (Bindi et al. 2000).

The great advantage of this corpus is that it is syntactically annotated with a shallow parsing technique. Texts are segmented into minimal structured units (*chunks*), on the basis of syntactic information (see Lenci et al. In print for details). Fig. 1 below shows a simplified version of a chunked sentence:

Il piano di risanamento

```
[[[chunk type: Nominal]
     [determiner: IL: definite article, masculine
     singular]
     [agreement: masculine singular]
     [potential governor: PIANO# noun, masculine
     singular]
 [[chunk type: Prepositional]
     [preposition: DI, preposition]
     [agreement: masculine singular]
     [potential governor: RISANAMENTO# noun masculine
     singular]]]
```

Fig. 1 Example of a chunked text

The resulting sequence of chunks constitutes the input of a dependency parser, which establishes dependency relations between the chunks (Bartolini 2004). Because the syntactic relation that links the head noun and the PP in a N+PP ICN is a modification relation, the use of a syntactically analyzed corpus and of a dependency parser is an optimal choice to eliminate from the dataset of ICN candidates many of the similar but not relevant syntactic structures present in the corpus.

First step: syntactic pattern extraction

The first step was to acquire all N+PP occurrences satisfying certain mor-

pho-syntactic conditions, exploiting the potential of our syntactically annotated corpus.

A special set of rules has been designed to extract only the following three basic patterns:

– "noun + "*di*" + noun";
– "noun + "*da*" + noun";
– "noun + "*a*" + noun".

The rules involve a nominal or an adjectival chunk in the first position followed by a prepositional chunk in second position, with various morphological constraints on the nouns functioning as potential governors and on the type of article inside the prepositional chunk. In order to reduce the noise in the output as much as possible, pronouns, proper names, articles and numerals were discarded. Numerals were left out in order to avoid cases like "una delle ragazze non era presente alla festa" ("one of the girls wasn't present at the party"), which surely are not instances of ICNs. The rules are written so as to specify only the particular preposition that has to be present in the target expressions, that is *di*, *da* or *a*. In this way, a dependency relation between the nouns of the two chunks isestablished in a way that satisfies the required conditions, and a label is attached in order to distinguish the structural types of CN (i.e. N_*di*_N, N_*a*_N, or N_*da*_N).

Ex. (7) is an example of the output of a rule; the label attached to the left indicated the structural type of ICN:

Ex. (7)

N_DI_N (SCATOLA,CARTONE <intro = DI>, Determiner = Absent)

As can be seen, the rules recognise and distinguish also the type of article within the PP: whether it is absent, definite or indefinite.

All this data have been imported into a database that stores all relevant syntactic features in distinct fields. Additionally, a field with the frequency of occurrence of all expressions grouped by syntactic type have been added. An example of the dataset containing syntactic information is given in Table 1.

Table 1
Example of the syntactic dataset

HeadNoun	Modifier	Determiner	Syntactic Structure	Syntactic type Frequency
PADRONE	CASA	0	N_DI_N	37
CONSIGLIO	STATO	0	N_DI_N	33

Second step: augmenting the data with semantic information

The main claim of our investigation is that Italian productive N+PP CNs are semantically motivated (constructions): their (implicit) semantic relation is a function of the interaction between the semantics of both head and modifier nouns. Therefore, we exploit a semantic lexical database to augment the syntactically based dataset described above with information on the semantic class of the nouns. The Lexical database in question is the SIMPLE-CLIPS Lexicon for Italian (Lenci et al. 2000, and Ruimy et al. 2002). This lexicon has been chosen mainly because it integrates ontological information with articulated lexical representations based on QS.

In the following section we briefly describe the structure of the SIMPLE lexicon.

The SIMPLE-CLIPS Semantic Lexicon

SIMPLE[11] was a lexicographic project for the construction of harmonized semantic lexicons for 12 European languages[12]. Such lexicons were designed to be closer to semantic-conceptual resources than to traditional lexicons, and application independent (Lenci et al 2000). SIMPLE-CLIPS has, moreover, an empirical basis, in that the senses encoded come from a set of harmonized and representative corpora for the 12 languages involved[13]. The multilingual aspect of the project has determined the need of identifying elements of the semantic vocabulary that could be used to express meaning components of the senses encoded in such a way as to keep the lexicon both language independent and capable of capturing those generalisations that are useful for different NLP needs.

To satisfy these needs the SIMPLE model is based on GL theory, which, especially in the Qualia structure, is able to capture the various dimensions of

[11] Acronym for *Semantic Information for Multipurpose Plurilingual Lexicons*; a Language Engineering project funded by the Eurpean DG-XIII. This project has been continued at the national level with the CLIPS project.

[12] Catalan, Danish, Dutch, English, Finish, French, German, Italian, Modern Greek, Portuguese, Spanish, and Swedish.

[13] These corpora were built within a previous project, called PAROLE. In the following we will refer to the PAROLE corpus intending, for reasons of simplicity, the corpus of Contemporary Italian built within that project and on which the lexicon is based.

word meaning, not only the taxonomic one that is dominant in almost any other computational lexicon, i.e. WordNet. The Qualia Structure, thus, constitutes the basic syntax for the construction of word meaning.

SIMPLE lexicons are based on three formal entities: Semantic Units, Semantic Types and Templates. Word senses are encoded as Semantic Units (Usem); each Usem is assigned a Semantic Type and other types of information that are specified in the Template associated with it[14].

For example the Usem for *coltello* 'knife' will be represented as in Figure 2:

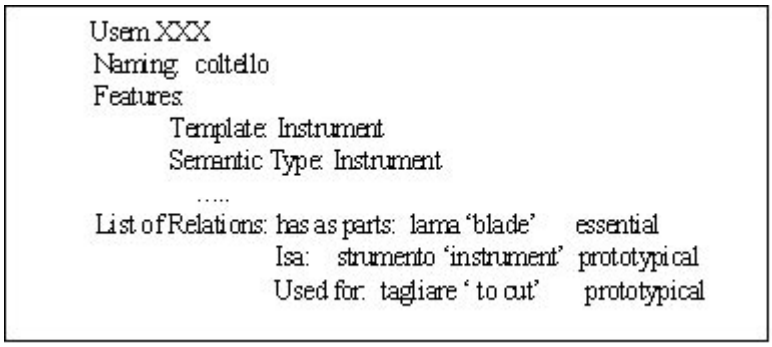

Fig. 2: SIMPLE-CLIPS Entry for *coltello* 'knife'

The set of Usem constitutes the lexicon for a given language, whereas the set of semantic types constitutes the Ontology, or the conceptual nucleus shared by all 12 lexicons.

General Architecture

SIMPLE lexicon is composed mainly of three integrated parts: 1- the Semantic Units: i.e. one entry for each sense of a lemma; 2- a semantic type hierarchy (or Ontology); 3- a semantic relation hierarchy: the Extended Qualia Structure. All three parts are linked to one another, and all of them contribute to the specification and encoding of semantic units. Each semantic unit is

[14] A Template is a representational tool containing predefined elements and slots for given semantic classes.

linked to the ontology through its template type specification, and is linked to other senses in the lexicon through qualia relations.

The Ontology

SIMPLE Core ontology is based on EuroWordNet base concepts and is constituted by those semantic types (concepts) that have been identified as central and common to all the 12 lexicons for the languages involved in the project[15]. These types represent the highest nodes in the hierarchy.

The Core Ontology has been constructed also taking into account the principles of qualia structure, in such a way as to allow also for a horizontal organization of semantic types. The idea that lies behind orthogonal architectures seems to be a good approach to overcome the limits of conventional type systems, structured according to the taxonomic principle, which is based exclusively on the ISA relation. In SIMPLE-CLIPS, and in GL, such relation is encoded in the formal Quale, and it bears the responsibility of the "vertical" organization of the ontology and of the lexicon. The other three qualia, conversely, specify semantic aspects and relations of the "horizontal" dimensions of word senses. The combination of the vertical and the horizontal dimension transforms the tree-structured ontology into a lattice hierarchical structure. The basic and highest nodes in the hierarchy are four, namely: [entity], [telic], [constitutive], [agentive]. As we can see, three of them are also those qualia that represent the horizontal dimensions of word meaning, and are also the highest nodes of the Extended Qualia Structure.

The Extended Qualia Structure

As mentioned above, the Qualia Structure is one of the most interesting aspects of the Generative Lexicon Theory (Pustejovsky, 1995), in that it decomposes the internal constitution of lexical items into 4 basic roles (Formal, Constitutive, Telic, Agentive), thus allowing to systematically structure and specify the relationships among lexical items both paradigmatically and syn-

[15] This way the multilingual aspect of the lexicons is guaranteed; actually, a cross-language link can be established automatically by the linking of all lexicons to the already existing Interlingual Index of EuroWordNet (P.Vossen 1998).

tagmatically. The SIMPLE-CLIPS lexicon implements this structure and it further specifies, for each role, more specific relations, encoded as a hierarchical structure, that is the Extended Qualia Structure (hereafter EQS).

The four qualia roles have been extended and implemented as relations between senses; each role represents the highest node of a hierarchy of more specific relations that extend the meaning of the four basic roles. It is argued that in this way finer distinctions among senses and semantic types can be better expressed. In Table 2 we give a small sample of the extended relations and of the hierarchical structure.

Table 2
The EQS

Relation	Class
Agentive	Agentive
Causedby	Agentive
Hasaspart	Constitutive
TypicalLocation	Constitutive
Madeof	Constitutive
Usedfor	Telic
Objectoftheactivity	Telic

In the present experiment we used EQS as a repository of useful relations for the representation of ICNs. Now, the objective is to find patterns of ICN formation on the basis of their syntactic structure and semantic/ontological information of their elements, so as to assign them (possibly only) one semantic relation taken from our predefined set, the EQS. The advantage of using these relations is that ICNs, once identified and annotated, would be automatically linked to other elements in the lexicon.

Linking Syntactic and Semantic Information

The data, annotated with the specific dependency relation and obtained by the preprocessing phase previously described, is integrated into a light version of the lexical semantic DB described above, so that each noun is linked to its corresponding semantic unit. A semantic unit in our version of the DB

contains only information about its semantic and template type. In this way we augment the original data set, which contains only syntactic information, with the semantic class of head and modifiers nouns.

An example of the augmented dataset is given in table 3:

Table 3
Augmented dataset

BORSA	Container	CUOIO	Artifactual_material	0	N_DI_N	1	29
BOTTIGLIA	Amount	ACQUA	D_3_Location	0	N_DI_N	3	21
BOTTIGLIA	Container	ACQUA	Natural_substance	0	N_DI_N	3	38

It is worthy of note that, due to polysemy and because no Word Sense Disambiguator had applied before, one syntactic unit (word) may be assigned to more than one semantic class, as one can observe in the last two rows in fig. 5. Therefore, the number of items in the DB, once semantic information is added, greatly increases. This creates ambiguity in the following step, which represents and attempts to uncover regular patterns of ICN formation.

Discovering Semantic Patterns

What has been described so far are the automatic steps performed to obtain good candidates of Italian N+PP CNs, annotated with semantic information.

The following step consists in the detection of productive syntactic-semantic patterns that allow us not only to establish paradigms of ICN formation, but also to assign to each token the corresponding implicit semantic relation. This phase has been carried out manually, after an automatic ranking of candidates based on raw frequency (see below for details), on the basis of the intuition of a native speaker of Italian.

From the dataset described above, we have selected those types of N+PP whose heads are related to Semantic Units (Usems) belonging to the following template types in the lexicon: Artifact, Instrument, Container and Vehicle. This sample contains 3453 different types of candidate ICNs. A restriction on the frequency of the syntactic types has been set to those types that show an intermediate frequency. In this way we reduce the number of Named Entities and fully lexicalised expressions (which usually have high frequencies), and at the same time avoid extracting fully regular syntactic phrases (and erro-

neous expressions). This step significantly narrows down the dimension of the data that has to be checked manually in order to identify the potential productive syntactic-semantic constructions. With this procedure we identified a number of patterns that seemed to be productive.

Additionally, our findings seem to support the claim that productive patterns often function as a basis for metaphorical extension, giving rise to lexicalised, or idiomatic CNs. Under the Artifact/Container*/Instrument/Vehicle + Natural_Substance pattern (in Table 4), for example, we find the idiomatic expression *maschera di sangue* (literally 'mask of blood') which can be intended to be 'a mask made of blood', at a very simple level of interpretation.

A set of the discovered patterns is presented in Table 4 below.

Table 4
Productive patterns identified

Semantic Patterns	**Syntactic Relation**	**Example**	**Semantic Relation**
Artifact+Location	N_da_N	*Maschera da teatro*	Typically used in
Artifact/Container*/Instrument /Vehicle+Natural_Substance	N_di_N	*Scatola di ferro*	Made of.
Artifact+Part	N_a_N	*Tavola a vela, veicolo a ruote*	Has as parts
Artifact/Container*/Instrument /Vehicle +Substance	N_di_N	*Coppa di veleno, maschera di lattice*	ambiguous
Artifact+Substance	N_a_N	*Maschera a gas*	Constitutive
Artifact/Container*/Instrument /Vehicle+Artifactual_Material	N_di_N	*Foglio di carta*	Made of

Such manually identified patterns have been subsequently exploited to retrieve those ICNs that instantiate the specific pattern, independently of their frequency. In this way we avoided the problem of data sparseness. In fact, applying these patterns to the semantically augmented tokens dataset, we are able to retrieve also hapax legomena (syntactic-semantic patterns with frequency equal to 1), which constitute 80% of our data, and to annotate them with the appropriate semantic relation. See Table 5 as an example.

What use can be made of such patterns? The "regular" patterns identified could be added to lexicon as schemas, or be used to retrieve and encode ICNs, at least by automatically specifying the appropriate semantic relation.

Table 5
Madeof [Artifact][artifactual_material]

Head	**Usem_template**	**Modifier**	**Usem_1_template**	**Det**	**SyntRel**	**Relation**
FOGLIO	Artifact	CARTA	Artifactual_material	0	N_DI_N	Madeof
PLACCA	Artifact	METALLO	Artifactual_material	0	N_DI_N	Madeof
FOGLIO	Artifact	PLASTICA	Artifactual_material	0	N_DI_N	Madeof
RECINZIONE	Artifact	LAMIERA	Artifactual_material	0	N_DI_N	Madeof

Problematic Issues

Working with real linguistic data, as we have done, shows that there are still many difficulties and controversial points. First of all, it is often difficult, even for the native speaker, to decide which relation, among those available in the EQS, is most suitable; a problem that has been already highlighted in our theoretical investigation of Qualia representation above. As we can see in Table 4, for the ICN *maschera a gas*, for example, we was not able to find a specific relation among the given ones, so that we indicated generally which role is involved, i.e. the constitutive role, one of the highest nodes in the EQS. This may well be a limit of the present investigation or of the EQS; however, it might also be a limit of the model itself: the qualia structure, in fact, seems to be too rigid to account satisfactorily for all semantic nuances that lexical items can assume in language use.

Another problem that we encountered is the inherent ambiguity of some patterns. Generative Lexicon Theory itself argues that ambiguity cannot be solved within the lexicon, but that it can be neutralized only once the context is taken into account.

Let us consider for example the following patterns identified in our experiment (Table 6):

Table 6
An example of ambiguity between patterns

Container + Natural_Substance	N_di_N	Madeof
Container + Natural_Substance	N_di_N	Contains

These seem to be cases of systematic polysemy, which is related, in fact, to the nature of the senses of the items involved. Such ambiguous tokens pre-

sented difficulties also to the human reader: a container may be either made of a natural substance or may contain it, which depends both on the nature of the container and of the specific substance (i.e. whether it can be used as a material or not).

Finally, one must take into consideration also another type of ambiguity in the data: the same lemma may belong to different semantic classes. An approach like the one we have devised must necessarily work in conjunction with a word sense disambiguator, in order to select for the proper senses of lemmas in specific ICNs, or to rely on human disambiguation, as we have done in the present experiment.

Finally, one consideration must be made about the limits of an approach like the one adopted here. In order to discover semantic paradigms of ICN formation we used semantic/ ontological information. Theoretically this is an interesting and motivated choice; from a practical point of view, however, the use of an ontology is problematic. A human made ontology is always a partial "picture" of the world and must face practical needs: it must fit certain dimensions, it must suit the tasks it has been designed for, and so on. General-purpose ontologies, additionally, are often too general for real system to use. Therefore, the semantic types attributed to each sense of a lemma may not be specific enough for all possible contexts.

Conclusions

The present contribution has addressed the issue of ICNs from an (automatic) acquisition and representational perspective, especially for NLP. We observed that, just like English noun compounds, ICNs blur the distinction between the syntactic and the lexical component because they are (at least) partially non-transparent but, nevertheless, show regularities both at the syntactic and at the semantic level. Therefore, we have referred to a non-traditional generative theory of lexicon, namely Generative Lexicon as a model for the representation/ interpretation of ICNs. Starting from an experiment by Johnston and Busa (1999) on the interpretation of English compounds and their Italian equivalents, we explored the representational power of qualia structure with respect to ICNs. We found that in many cases Qualia Structure is expressive enough to represent and ICNs in the lexicon, also for translation purposes, provided that the same entity is profiled lexically in the same way. However, there are cases where things get more complicated and qualia representation appears to be too rigid: we discussed, for example, cases of ambiguity of

the qualia role involved. Nevertheless, the retrieval of the appropriate semantic relation is a fundamental step for the interpretation, representation or annotation of ICNs. Once the specific semantic relation is identified, it can be used to specify the relative qualia role in the qualia structure of the construction, following a predefined schema. Otherwise, it can be used to specify patterns of ICN formation, which, in their turn, would be employed for the recognition and annotation of ICN instances in texts, or for their online interpretation.

The second and central part of our work consisted of an experiment for the (semi) automatic detection of syntactic-semantic patterns of ICNs, aimed specifically at the identification of the appropriate semantic relation. Using a syntactically annotated corpus as input, a dependency syntactic parser for Italian, and a lexical semantic DB, we obtained, through a series of subsequent steps, a dataset that contains all good candidate ICNs annotated with information on the syntactic and semantic patterns they instantiate. This dataset has proved useful to detect productive patterns of ICN formation, and to determine the specific semantic relations. The syntactic-semantic patterns identified have subsequently been used to retrieve low frequency occurrences of the same pattern, and then to annotate automatically each occurrence with the corresponding semantic relation.

However interesting the results may be, our approach encountered some difficulties as well. The first problem that emerged is related to the rigidity of Qualia Structure and to the limitations deriving from postulating a limited predefined set of relations. Secondly, a major difficulty was created by the use of a semantic type system (or ontology), which cannot be equally detailed and adaptable for all possible domains and contexts. Finally, problems arose because of polysemous items, which make the number of syntactic-semantic patterns, represented as rows in the DB, explode, create ambiguity among patterns. While the first two problems do not seem to be easily solvable with our approach, the third one could be avoided making use of a word sense disambiguator. This possibility, however, is left unexplored, and may be the topic of further work.

References

Bartolini, Roberto et al. (2004) "Semantic Mark-up of Italian Legal Texts Through NLP-based Techniques". In Proceedings of the IV International Conference On Language Resources Evaluation (LREC2004), 795- 798.

Bindi, Remo et al. (2000) *PAROLE-Sottoinsieme*. ILC-CNR Internal Report, Pisa: ILC.

Busa, Federica. (1996) *Compositionality and the Semantics of Nomin*als. PhD Thesis. Brandeis (MA): Brandeis University.

Downing, Pamela (1977). "On the creation and Use of English Compound Nouns". *Language*, 53, 810-842.

Giovanardi, Claudio (2004) "Chimica" In M. Grossmann and F. Rainer (eds.), 580-584.

Grossmann, Maria and Franz Rainer (eds.) (2004) *La Formazione Delle Parole In Italiano*. Tuebingen: Max Niemeyer Verlag, 56-68.

Johnston, Michael, Busa, Federica. "Qualia structures and the compositional interpretation of compound". In E. Viegas (ed.) (1999) *Breadth and depth of semantic lexicons*. Dordrecht: Kluwer Academic Publishers.

Lenci, Alessandro et al. (In print) "CHUNK-IT. An Italian Shallow Parser for Robust Syntactic Annotation", *Linguistica Computazionale*, (TR16\CNR-ILC\2000).

Lenci Alessandro et al. (2000) "SIMPLE: A General Framework for the Development of Multilingual Lexicons", *International Journal of Lexicography*, XIII (4), 249-263.

Levi, Judith N. (1978) *The syntax and semantics of Complex Nominals*. New York: Academic Press.

Lyons, John (1977) "The Lexicon". In Lyons, J. *Semantics*. Cambridge: CUP, 512-569.

Petrocelli, Simonetta. (1992) "La Composizione Nominale in Italiano e in Tedesco" in *Studi di Linguistica Italiana Teorica e Applicata*. XXI, 65-82.

Pustejovsky, James (1995) *The Generative Lexicon*, Cambridge, MA, The MIT Press.

Quochi, Valeria (2004) "Representing Italian Complex Nominals: A Pilot Study". In *Proceedings of the IV International Conference On Language Resources Evaluation* (LREC2004),Volume IV, 1863-1866.

Ruimy Nilda et al. (2002) "CLIPS, A Multi-level Italian Computational Lexicon: a Glimpse to Data". In *Proceedings of the III International Conference On Language Resources Evaluation* (LREC2002), Volume III, 792-799.

Voghera, Miriam. (2004) "Polirematiche" In Grossmann, M., Rainer. F., (eds.) 56-68.

Vossen, Piek at al. (1998) "The EuroWordNet Base Concepts and Top Ontology" Deliverable D012, D034, D036, LE2-4003, http://www.let.uva.nl/ewn.

Warren, Beatrice. (?1978) *Semantic Pattern of Noun-Noun Compounds*. Acta Gothenburgensis. Studies in English. Gotheborg: University of Gotheborg.

Weinrich, Harald (1977) "L'Antropologia delle Preposizioni Italiane." *Studi di Grammatica Italiana*. Vol. VII, Firenze: Accademia della Crusca, 255-280.